Modeling Continuous Data with Discrete Bins

About the Relative Speed of Processes

Daniel W. Heck

STATISTICAL MODELING in PSYCHOLOGY

FREIBURG HEIDELBERG LANDAU MANNHEIM TÜBINGEN

2019-04-30

MPT Modeling with Continuous Data

1) MPT-RT: Modeling response times with histograms
   - Heck & Erdfelder (2016)
2) GPT (generalized processing tree): Parametric modeling
   - Heck, Erdfelder, & Kieslich (in press)
3) RT-MPT: Serial-process model for response times
   - Klauer & Kellen (2018)

MPT Models and Continuous Variables

- Discrete-state modeling for discrete and continuous variables
  - Response times, confidence ratings
  - Process tracing measures (eye or mouse tracking)
  - Neurophysiological data (e.g., amplitudes of ERP signals)
- Structure of the data:
  - In each trial we observe one discrete response and one or more continuous values (e.g., response time)
  - In standard MPT modeling, we would simply ignore all continuous measures and make a frequency table of discrete responses

Item Type	Discrete Response	Response time
Target	"old"	930
Target	"new"	1532
Target	"old"	1240
...	...	
Lure	"old"	798
Lure	"new"	2332
...	...	

Mixture distribution

- All MPT extensions assume mixture distributions for discrete and continuous observations
  - A) Latent RTs: Different processing branches of the MPT model result in different latent distributions $g_j(t)$
  - B) Observed RTs: A mixture distribution, defined as $f(t) = \sum_j p_j g_j(t)$
  - C) The mixture weights p_j are determined by the MPT structure (= branch probabilities)

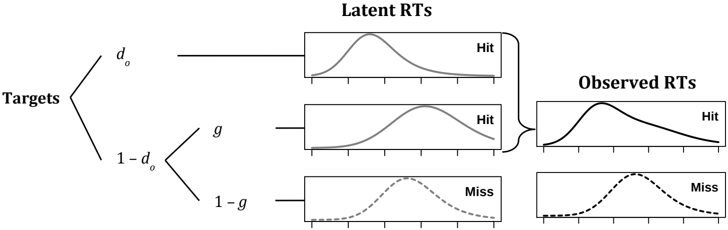
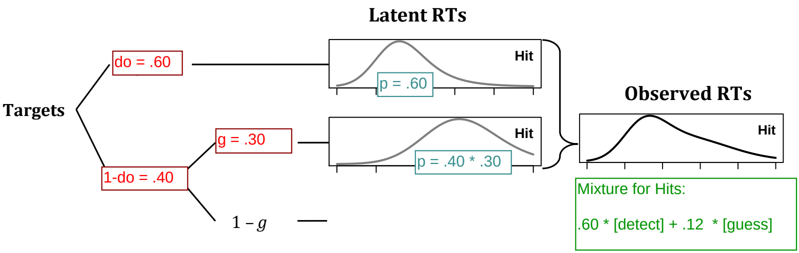


Illustration: 2-high threshold model

- Latent RTs:
  - $g_{\text{detect}}(t) =$ RT distribution for detection
  - $g_{\text{guess}}(t) =$ RT distribution for guessing
- Observed RTs for correct "old" responses to targets:

$$f(t, \text{Hit}) = d_o \cdot g_{\text{detect}}(t) + [(1 - d_o)g] \cdot g_{\text{guess}}(t)$$



Three different approaches

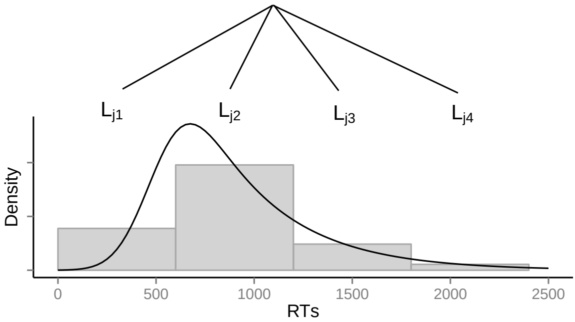
- The three methods *all* assume mixture distributions for continuous variables
- Main difference: Assumptions for the component distributions $g_j(t)$
  - A) Histogram/nonparametric
  - B) Any parametric distribution
  - C) Serial-processing assumptions

MPT-RT: Modeling response times with histograms

- Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review, 23*, 1440–1465.
- Heck, D. W., & Erdfelder, E. (2017). Linking process and measurement models of recognition-based decisions. *Psychological Review, 124,* 442–471.

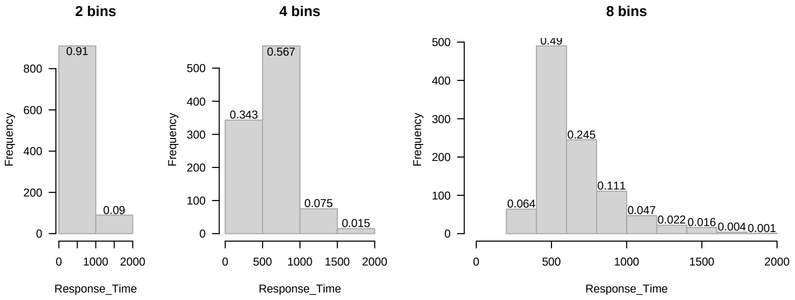
Histogram-Based Approach (Heck & Erdfelder, 2016)

- Categorize RTs into discrete bins (Yantis, Meyer, & Smith, 1991)
  - Example: “Very fast”, “fast”, “slow”, “very slow”
- State-specific distributions are modeled by the parameters L_{jb}:
  - $L_{jb} =$ height of the histogram bins
  - $L_{jb} =$ probability that state j results in observation in the b-th interval



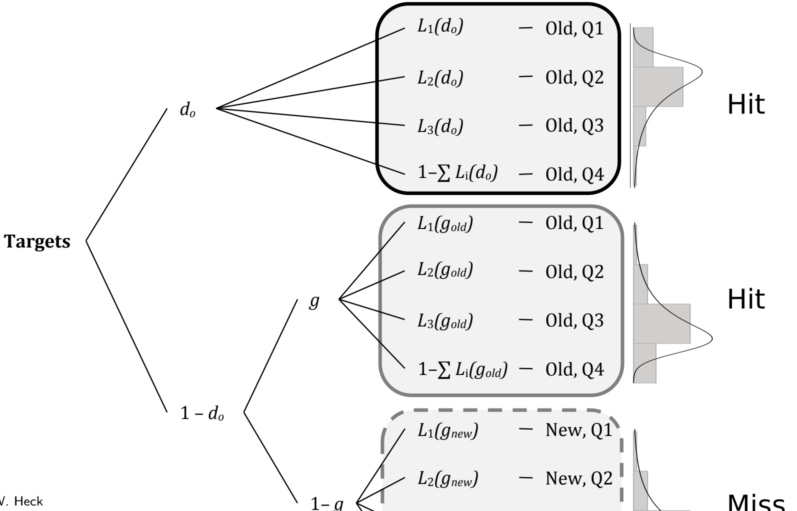
Categorizing RTs

- Depending on the research question and the number of observations, we can use more or less bins
- Note that the bin probabilities always must sum to one!
  - Hence, for 2 bins, we need 1 L-parameter
  - Hence, for 8 bins, we need 7 L-parameters



The RT-extension results in a new (larger) MPT model

- Each set of L parameters represents a histogram for *one* latent RT distribution



Using Histogram-MPTs in Practice

1. Categorize continuous variable into discrete bins
2. Derive constraints which of the latent component distributions are identical
   - Example: Identical RT distribution of "guessing old" for targets and lures
3. Fit the new RT-MPT model
   - Data: Frequencies for all combinations of discrete responses and RT bins

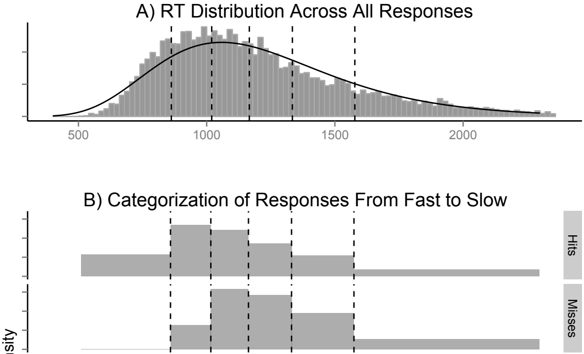
MPT category	RT bins			
	Very fast	Fast	Slow	Very Slow
Target: Hit	44	36	15	4
Target: Miss	15	8	13	23
Lure: False alarm	4	17	22	19
Lure: Correct rejection	31	41	9	4

Details

- Concerning (1): How to define RT boundaries for the bins?
- Concerning (2): Is the new model identifiable?

A Principled Strategy to Categorize RTs (details: Heck & Erdfelder, 2016)

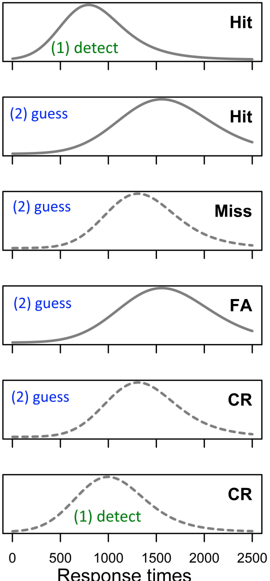
- Compute separate RT bounds per participant (individual differences)
  - Interpretation: "Are responses fast or slow *relative* to the overall speed of responding of a person?"
- With 2 RT bins:
  1) Compute the geometric mean across *all* RTs: `bound = exp(mean(log(RTs)))`
  2) Categorize responses as "fast" or "slow"
- Illustration for more RT bins:



Problem 2: Identifiability

Identifiability of latent RT distributions

- The number of latent RT distributions must be equal or smaller than the number of observed distributions
  - 2HTM: Maximum of 4 latent RT distributions (observed: hit, miss, FA, CR)
- Some latent distributions directly result in observable distributions
  - These latent distributions are directly identifiable
  - 2HTM: Misses and FAs are always guessing RTs!
- A stepwise procedure allows to check the identifiability of the remaining component distributions
  - cf. Appendix



Summary

Recipe for MPT-RTs in practice

1. Categorize continuous variable into discrete bins
   - Example: RTs faster or slower than geometric mean?
2. Derive constraints which of the latent component distributions are identical
3. Check identifiability (and revise model)
4. Collect data with RTs
5. Fit the new MPT-RT model
6. Test hypotheses about the relative speed of processes (L parameters)
   - Very simple for 2 RT bins: one L parameter per process ("fast" vs. "slow")
   - Are processes equally fast? (equality constraints: $L_f = L_s$)
   - Is process i faster than process j? (order constraints: $L_f > L_s$)

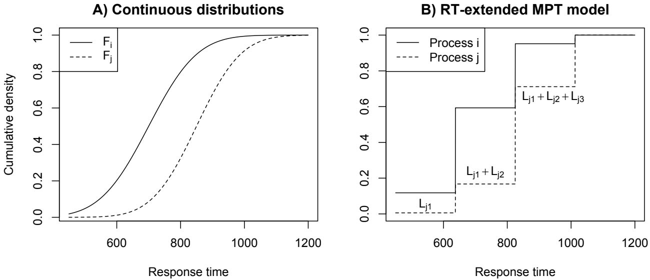
Summary

Advantages of the Histogram Approach

- Does not require parametric assumptions for latent distributions
- Simple: one can use standard MPT software
  - multiTree, TreeBUGS
- Allows novel tests of theories by including RTs
  - Recognition heuristic (Heck & Erdfelder, 2017)

Appendix

- Is process i faster than process j? (order constraints)



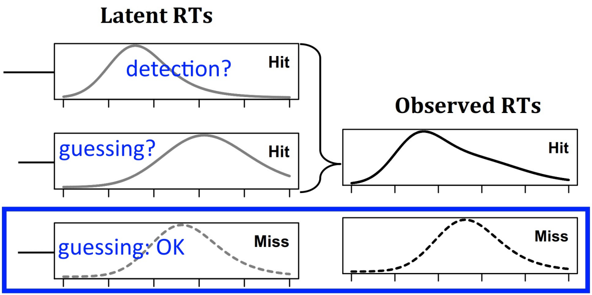
A stepwise procedure to check identifiability

1) Those latent distributions that directly result in observable distributions are directly identifiable
2) Look for 2-component mixtures
3) Is one of latent RT distributions directly identifiable from the first step?
4) It follows that the second RT distribution is also identifiable!
5) Look for 3-component mixtures
6) Check whether 2 of the 3 components are identified
7) ...

Details: Next slides

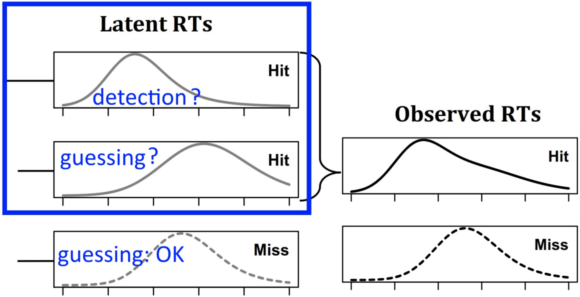
Step 1:
Observable latent RT distributions

2HTM: guessing RTs = Miss RTs



Step 2:
Check 2-component mixtures

2HTM: Hit RTs



Step 3: Check whether components are identifiable

2HTM: detection RT identifiable